Department cop

CONSISTENCY OF SMOOTHING WITH RUNNING LINEAR FITS

Art B. Owen

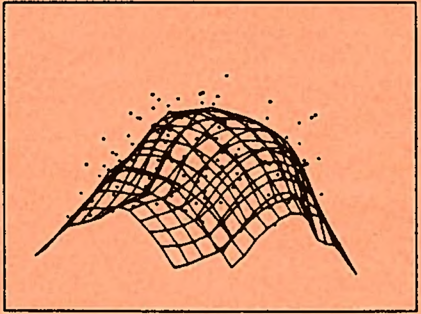
and

Joseph C. Marhoul

Technical Report No. 8

November 1984

Laboratory for Computational Statistics



Department of Statistics
Stanford University

This document and the material and data contained therein, was developed under sponsorship of the United States Government. Neither the United States nor the Department of Energy, nor the Office of Naval Research, nor the U.S. Army Research Office, nor the Leland Stanford Junior University, nor their employees, nor their respective contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any liability or responsibility for accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use will not infringe privately-owned rights. Mention of any product, its manufacturer, or suppliers shall not, nor is it intended to, imply approval, disapproval, or fitness for any particular use. A royalty-free, nonexclusive right to use and disseminate same for any purpose whatsoever, is expressly reserved to the United States and the University.

CONSISTENCY OF SMOOTHING WITH RUNNING LINEAR FITS

Joseph C. Marhoul

and

Art B. Owen

Department of Statistics, Stanford University

and

Stanford Linear Accelerator Center

L.C.S. Technical Report No. 8

November 1984

Keywords: Smoothing, Nonparametric Regression, Consistency

Abstract

We establish the mean square consistency of running (ordinary) least squares linear regression smoothers, under realistic conditions on the joint distribution of the abscissa and ordinate (X and Y below) variables. The windows used in the running least squares fits need not be centered on the points for which they are used. In fact, we show that taking a window of points entirely to one side of a data point, fitting a line to that window and using the value of that line at the target point is consistent. It follows that the Supersmoother of Friedman and Stuetzle (1982) and the Split Linear Smoother of McDonald and Owen (1984) are both consistent.

This work supported by an N.S.E.R.C. Postgraduate Scholarship, Office of Naval Research contract N00014-83-K-0472, Office of Naval Research grant N00014-83-G-0121, and U.S. Army Research Office under contract DAAG29-82-K-0056.

1. Running Linear Smoothing.

Given observations $(X_i, Y_i) \in \mathcal{R}^2, 1 \le i \le n$ a running linear smooth value at x is $\alpha + \beta x$ where α and β are regression coefficients from a linear regression of Y on X over a set of observations indexed by $J(x)$.

In practice $J(X_i)$ typically consists of the union of: the indices of the smallest $k_n/2$ points in $\{X_j : X_j > X_i\}$, the indices of the largest $k_n/2$ points in $\{X_j : X_j < X_i\}$ and $\{i\}$ itself, with sensible modifications to handle ties and end effects. Furthermore $J(x)$ is not usually calculated for x's that do not correspond to sample points, it being more expedient to interpolate if necessary. In this paper it is more convenient to define the smoother at all points without resort to interpolation. The above describes a central running linear smoother, the adjective 'central' serving to distinguish it from one sided smoothers in which $J(x)$ consists of the k_n nearest neighbors of x on the left (or right).

The Supersmoother of Friedman and Stuetzle (1982) combines several central smoothers, differing only in the value of k_n. Its design goal is to make more use of the smaller windows in regions of $X-$space where the curvature of the regression of Y on X seems large relative to the variance of Y and to emphasize the larger windows where the curvature is smaller, so as to locally trade off bias versus variance.

The Split Linear Smoother of McDonald and Owen (1984) combines central smoothers with left and right sided smoothers. The design goal is to provide an edge-detecting smoother that produces output that is piece-wise smooth with a small number (possibly zero) of discontinuities in the curve or its first derivative. It does this by taking a weighted average of the smooths at each point; near a discontinuity it uses larger weights for the windows that extend in the direction opposite the discontinuity.

Because these smoothers are used as building blocks in non-parametric regression techniques such as projection pursuit regression (Friedman and Stuetzle (1981)) and A.C.E. (Breiman and Friedman (1984)) proofs of their consistency have ramifications beyond smoothing.

Stone (1977) shows that linear fits over sets of nearest neighbors are consistent when trimmed. The nearest neighbor linear fits can be expressed as a weighted average of the Y values in the neighbor set. Trimming involves adjusting those weights if necessary to make sure that their ratios to the weights of some consistent estimator are uniformly bounded above and below. The consistent estimator may be taken to be a nearest neighbor average. He states that linear fits to nearest neighbors are not necessarily consistent.

Breiman and Friedman (1982) show that a modified central running linear smoother is consistent. The modification is greatest at the end points of the sample where a lack of observations makes it impossible to form the usual symmetric nearest neighbor window. Unfortunately, the main reason for using linear as opposed to constant fits is to reduce bias at the ends. Their modification would be severe for a one sided smoother that effectively treats every point as an endpoint. (There are reasons other than consistency for making the modification. They need a smoother that is a bounded linear operator on the observed Y values whatever the X values, and the bound must be uniform in the sample size. Running linear smooths (central or sided as defined here) have bounds that increase as the square root of the window size.) Rather than changing the definition of the running linear smoother we place additional restrictions on the distribution of the observations. Fortunately the restrictions are realistic for applications.

By restricting the distributions we do not establish what Stone (1977) calls universal consistency. He also obtains L^r consistency for all $r \geq 1$ such that the r^{th} moment of Y is finite, whereas we only consider L^2.

2. Notation.

This section introduces the notation and defines a left sided running linear smoother.

The observations are a (finite prefix of) an infinite sequence of i.i.d. random variables (X_i, Y_i), $1 \leq i < \infty$. X and Y represent the complete sequences. The r^{th} order statistic among the first n terms of X will be denoted $X_{(r)}^{(n)}$ and similarly for Y. The first n terms will be collectively denoted $X^{(n)}$.

For each $x \in \mathcal{R}$ and each positive integer n define $J_n(x)$, the n^{th} window about x as follows: if $x < X_{(1)}^{(n)}$ or $x > X_{(n)}^{(n)}$ then $J_n(x) = \emptyset$, if $x \in X^{(n)}$ then $J_n(x)$ consists of the greatest k_n terms of $X^{(n)}$ that are less than or equal to x (if there are not k_n such terms take all such terms and if a tie need be broken take the term(s) with smallest index), otherwise take the smallest term in $X^{(n)}$ that is greater than or equal to x and the $k_n - 1$ largest other terms that are less than or equal to x with the obvious handling of ties and shortfalls . For the typical point x, $J_n(x)$ consists of its nearest neighbor on the right and its $k_n - 1$ nearest neighbors on the left. For any observation X_i in $X^{(n)}$ the window $J_n(X_i)$ has no points strictly to the right of X_i. By construction, except in the case $J_n(x)$ is void

$$\min_{i \in J_n(x)} X_i \leq x \leq \max_{i \in J_n(x)} X_i.$$

The smooth value at x based on the first n observations is denoted $m_n(x, X, Y)$. As written, it depends on the whole sequences X and Y, but it will really only depend on the first n terms of them.

If $J_n(x)$ is empty, then take $m_n(x, X, Y) = 0$. Otherwise compute

$$\hat{\mu}_n(x) = \frac{1}{|J_n(x)|} \sum_{i \in J_n(x)} X_i$$

and

$$\hat{\sigma}_n^2(x) = \frac{1}{|J_n(x)|} \sum_{i \in J_n(x)} (X_i - \hat{\mu}_n(x))^2.$$

The left sided running linear smooth value is

$$m_n(x, X, Y) = \frac{1}{|J_n(x)|} \sum_{i \in J_n(x)} Y_i[1 + \tilde{x}\tilde{X}_i]$$

where for any w, $\tilde{w}$ is 0 if $\hat{\sigma}_n(x) = 0$ and otherwise

$$\tilde{w} = \frac{w - \hat{\mu}_n(x)}{\hat{\sigma}_n(x)}.$$

Below, the dependence upon x and n of J, $\hat{\sigma}$ and $\hat{\mu}$ is sometimes suppressed.

The following identity will often be convenient

$$\sum_{i \in J} (1 + \tilde{x}\tilde{X}_i)^2 = |J|(1 + \tilde{x}^2).$$

The quantity $\tilde{x}$ is the distance of the target point from the window mean expressed in window standard deviations. A large magnitude indicates that the target point is not well represented by the window set and this will be reflected in bias and variance expressions below. The construction of J gaurantees, by Chebychev's inequality, that $\tilde{x}^2 \leq k_n$. Breiman and Friedman's (1982) modified running linear smoother truncates $\tilde{x}$ to ± 1 when it exceeds 1 in absolute value. They also note that $|\tilde{x}| \leq 1$ whenever (in a central window) there are equal numbers of points greater than and less than x. The construction above is in effect truncating $\tilde{x}$ at $\pm\sqrt{k_n}$, which gives less control, but obviates the need to modify the smoother within the range of the observed X_i. For convenience of exposition the smoother is zero outside that range, although there would be no difficulty in extrapolating by extending the smooth values at the left- and right-most sample points to the left and right of the sample respectively. If x is an atom of $\mathcal{L}(X_1)$ then $\tilde{x}^2 \rightarrow 0$ almost surely. Otherwise, under reasonable sampling conditions Lemma 2 in section 5 shows that for a one sided smoother $\tilde{x}^2 \rightarrow 3$ in probability.

3. Main Result.

This section treats the pointwise mean square consistency of running linear smoothers, which means the L^2 convergence of $m_n(x, X, Y)$ to $m(x)$. Several technical Lemmas proved in section 5 are used. The following assumptions will be used to prove consistency of running linear smoothers:

(I) X_i are *iid* from distribution $F = F_d + F_{ac}$ where

a) F has support on $[0, 1]$.

b) F_{ac} has a continuous positive density on $[0, 1]$.

c) F_d has a finite number of jumps.

(II) Y_i are conditionally independent given X with $\mathrm{V}(Y_i|X) \leq \sigma^2$ for some $\sigma^2 < \infty$.

(III) If $m(x) = \mathrm{E}(Y|X = x)$ then for all but finitely many points x $\exists M(x) < \infty$ such that $|m(x) - m(x')| \leq M(x)\ |x - x'|$.

The first assumption constrains the distribution of the X's. For technical reasons the support of these random variables has to be compact and have an absolutely continuous part with density continuous and bounded away from zero. Jump points are allowed, although there can not be an accumulation point of the jumps of the distribution. The uniform bound on the conditional variance of the Y random variables given the X variable is weaker than the standard homoscedasiticy assumption but stronger than the assumption that $Y \in \mathrm{L}^2$. Condition (III) is unusual in that it allows simple jump discontinuities in the regression.

Theorem *If conditions (I) - (III) hold, $k_n \to \infty$, and $k_n^3/n^2 \to 0$ then $\{m_n\}$ is mean square consistent at x for almost all x.*

Proof: Define $S = \{x : \forall \delta > 0 \;\; \mathrm{P}(x \leq X < x + \delta) > 0, \mathrm{P}(x - \delta < X \leq x) > 0$, and either $\mathrm{P}(X = x) > 0$ or condition III holds $\}$. From the hypothesis it is easy to show that $\mathrm{P}(X \in S) = 1$. From the triangle inequality

$$\left\{\mathrm{E}\big(m(x) - m_n(x, X, Y)\big)^2\right\}^{\frac{1}{2}} \leq \left\{\mathrm{E}\big(m(x) - m_n(x, X, m(X))\big)^2\right\}^{\frac{1}{2}}$$

$$+ \left\{\mathrm{E}\big(m_n(x, X, m(X)) - m_n(x, X, Y)\big)^2\right\}^{\frac{1}{2}}, \qquad (1)$$

where the first term on the right hand side represents bias^2 and the second term represents variance. $m(X)$ is the sequence $m(X_i)$, $1 \leq i < \infty$. It is sufficient to show both terms converge to zero. First the variance term is considered, using $\mathrm{E}_X(\,\cdot\,)$ to denote conditional expectation given the X-sequence.

$$\mathrm{E}(\; (m_n(x, X, m(X)) - m_n(x, X, Y))^2 \;)$$

$$= \mathrm{E}(\; \frac{1}{k_n} \sum_{i \in J} (Y_i - m(X_i))(1 + \tilde{x}\tilde{X}_i) \;)^2$$

$$= \frac{1}{{k_n}^2} \mathrm{E}(\sum_{i \in J} \mathrm{E}_X(\; (Y_i - m(X_i))^2 \;)(1 + \tilde{x}\tilde{X}_i)^2$$

$$+ \sum_{i \neq k} \mathrm{E}_X(\; (Y_i - m(X_i))(Y_k - m(X_k)) \;)(1 + \tilde{x}\tilde{X}_i)(1 + \tilde{x}\tilde{X}_k) \;)$$

$$= \frac{1}{{k_n}^2} \mathrm{E}(\sum_{i \in J} \mathrm{E}_X(\; (Y_i - m(X_i))^2 \;)(1 + \tilde{x}\tilde{X}_i)^2 \;) \qquad (a)$$

$$\leq \mathrm{E}(\frac{\sigma^2}{{k_n}^2} \sum_{i \in J} (1 + \tilde{x}\tilde{X}_i)^2 \;) \qquad (b)$$

$$=\mathrm{E}\Big(\frac{\sigma^2|J_n|}{k_n} \; \frac{1+\tilde{x}^2}{k_n} \Big) \qquad (c)$$

Equation (a) is a consequence of the conditional independence of the Y_i's given the X_i's. (b) follows from hypothesis (II) and (c) from the identity given in the first section. From Lemma 2 it follows that the last line converges to zero by the dominated convergence theorem.

The bias^2 term is

$$\begin{aligned}
\mathrm{E}\big(\, (m(x) - m_n(x, X, m(X)))^2 \, \big) &= \mathrm{E}\Big(\frac{1}{k_n} \sum_{i\in J} [m(x) - m(X_i)][1 + \tilde{x}\tilde{X}_i] \, \Big)^2 \\
&\leq \mathrm{E}\Big(\frac{1}{{k_n}^2} \sum_{i\in J} (m(x) - m(X_i))^2 \sum_{i\in J} (1 + \tilde{x}\tilde{X}_i)^2 \, \Big) \qquad (a) \\
&= \mathrm{E}\Big(\frac{|J_n|}{k_n} \; \frac{1+\tilde{x}^2}{k_n} \sum_{i\in J} (m(x) - m(X_i))^2 \, \Big) \qquad (b) \\
&\leq 2\mathrm{E}\Big(\sum_{i\in J} (m(x) - m(X_i))^2 \, \Big) \qquad (c)
\end{aligned}$$

where (a) is a consequence of the Cauchy-Schwarz inequality, (b) is yet another application of the identity of the previous section, and (c) follows from Lemma 2. The last term converges to zero by Lemma 3.

Hence for all $x \in S$, the estimate is mean square consistent.

4. Further Results.

This section extends the above result to central and right sided smoothers, discuss consistency under design measures on X, rates of convergence and global L^2 convergence.

Corollary 1 Under the conditions of Theorem 1, right sided and central smoothers are pointwise mean square consistent.

Proof For right sided smoothers, the result follows by symmetry. For central smoothers define $J_n^L(x)$ as for a left sided smoother, $J_n^R(x)$ as for a right sided smoother and put $J_n(x) =$ $J_n^C(x) = J_n^L(x) \cup J_n^R(x)$. The quantity $\tilde{x}^2$ for the central smoother will be smaller than nine times the largest such quantity from the one sided smoothers, and $(x-\mu)^2$ will be no larger for the central smoother than the largest such value from the sided smoothers. (The first bound is obtained by straightforward calculation and is very conservative.) It follows from the bias and variance bounds above that the central smoother is mean square consistent at all $x \in S$.

Corollary 2 Under the conditions of Theorem 1, any smoother that at each point is a convex combination of central, left, and right sided smoothers is pointwise mean-square consistent. This includes, with appropriate window sizes, the Supersmoother and the Split Linear Smoother.

Proof Immediate.

Remark 1 Pointwise L^2 consistency implies pointwise convergence in probability, which when established for almost all points (i.e. all $x \in S$) implies convergence in probability of $\hat{m}_n(X_0)$ to $m(X_0)$ where X_0 is independent of X and has the same distribution as X_1.

Remark 2 Instead of observing i.i.d. pairs (X_i, Y_i), consider choosing X according to a design measure on the sequence and observing Y whose terms are conditionally independent given X, and satisfy the distributional assumptions as above. Then $\mathrm{E}(\ (1+\tilde{x}^2)/k_n\ ) \to 0$ is sufficient to guarantee that the variance term will vanish as $n \to \infty$. If there is a positive minimum conditional variance of Y given $X = x$, then $\mathrm{E}(\ (1+\tilde{x}^2)/k_n\ ) \to 0$ is also necessary for the variance term to vanish. Similarly the bias term can be controlled by design. If all the X_i are sufficiently uniformly spaced then $\tilde{x}^2$ will be bounded in x and n.

Theorem 2 (Global L^2 convergence) Under the conditions of Theorem 1, and if $|m(x) - m(x')| \leq M|x - x'|$ the central smoother satisfies $m_n(X_0, X, Y) \to m(X_0)$ in L^2 where X_0 is independent of X and has the same distribution as X_1.

Proof Let Z be an indicator variable which is 1 when $\tilde{X_0}^2 \leq 1$. Then the variance term is bounded by

$$\begin{aligned}
&\frac{2\sigma^2}{k_n} + 2\sigma^2 \mathrm{E}(\ 1 - Z\ ) \\
\leq &\frac{2\sigma^2}{k_n} + 2\sigma^2 (P(X_0 < X_{(k)}^{(n)}) + P(X_0 > X_{(n-k)}^{(n)})) \\
\leq &\frac{2\sigma^2}{k_n} + 4\sigma^2 \frac{k_n}{n} \\
\to &0,
\end{aligned}$$

(where $X_{(r)}^{(n)}$ is defined in section 2), and the bias^2 term is bounded by

$$\begin{aligned}
&2\frac{M^2}{\gamma^2}\mathrm{E}\Bigg(\sum_{i=1}^{|J_n^L(X_0)|}(U_{(i)}^{(n)})^2+\sum_{i=1}^{|J_n^R(X_0)|}(U_{(i)}^{(n)})^2\Bigg)\\
\leq &4\frac{M^2}{\gamma^2}\frac{k_n^2}{n^3}\\
\rightarrow &0,
\end{aligned}$$

where $U_{(i)}^{(n)}$ is the i^{th} order statistic out of n independent uniform random variables and γ is the same quantity found in the proofs of the Lemmas.

Remark 3 Notice that the bounds above imply the squared error can be made to converge at rate $n^{-2/3+\epsilon}$ by letting k_n grow at a rate slightly slower than $n^{2/3}$. The optimal global rate of convergence assuming one derivative is $n^{-2/3}$ (Stone (1982)). This proof also goes through for the sided smoothers except that Z must indicate that $\tilde{X}_0^2 < B$ for some $B > 3$. (Recall $\tilde{X}_0^2 \rightarrow 3$ in probability.) It can be shown that $P(\tilde{x}^2 > B) \rightarrow 0$ at least as fast as k_n^{-1} and so the same rates obtain for the sided smoothers as for central ones. However, the main use of sided windows is for situations in which it is suspected that there is a discontinuity.

Remark 4 Compared to linear fits over nearest neighbor windows, the central smoother is based on points farther away from the target. On the other hand, a linear fit over the k_n nearest neighbors puts no bound at all on $\tilde{x}^2$ since, in the worst case, all the neighbors can be in a cluster on one side of the target point. Using symmetric nearest neighborhoods $\tilde{x}^2$ is bounded by k_n for all points and by 1 for most points. (The former was handy in some dominated convergence arguments.) Thus in addition to being faster to compute, linear fits over symmetric nearest neighborhoods are safer than those over nearest neighborhoods.

Remark 5 The Split Linear Smoother and the Supersmoother were designed to meet specific finite sample goals. We believe those goals to be more important than asymptotic behavior. The point of this paper is to show that without any modification in the observed range of X the attainment of finite sample goals is not at undue asymptotic expense.

5. Proof of Lemmas.

Conditions stated in section 3 are assumed to hold. The main idea driving the following Lemmas is that under the distributional assumptions (I)-(III) the observations X_i "sufficiently close" to x behave like observations from a uniform distribution. Once this correspondence is established exact calculations can be made for the uniform variates in a way that gives bounds for the quantities of interest. Lemma 1 makes precise the sense in which the points of $J_n(x)$ get "sufficiently close" to x. Lemma 2 provides a construction that bounds the small order statistics of the $x - X$'s by the order statistics of a uniform distribution. This construction is then exploited by Lemmas 2 and 3.

By the definition of $J(x)$ there may exist an element of $J(x)$ to the right of x. In Lemmas 2 and 3 that element allows the application of Chebychev's inequality. However, explicit consideration of that point would add unnecessary complexity involving quantities of $O(1/k_n)$ to the following proofs. Therefore with the exception of the appeal to Chebychev's inequality that point is not considered. The reader is invited to make the necessary (minor) alterations to include the point if so motivated.

Lemma 1: $\forall x \in S$, $\max_{i \in J_n(x)} \; |x - X_i| \xrightarrow{\text{as}} 0$.

Proof: Let $T_n^m = \sum_{j=1}^n \chi\{x - 1/m < X_j \leq x\}$. From definition, $T_n^m > k_n$ implies $\max_{i \in J_n(x)} |x - X_i| < 1/m$. From the law of large numbers $\frac{1}{n} T_n^m \xrightarrow{\text{as}} \mathrm{P}(x - 1/m < X \leq x)$ which is positive by hypothesis. Since $\frac{k_n}{n} \to 0$, there exists a null set N_m such that on N_m^{C} lim sup $\{\max_{i \in J_n(x)} \; |x - X_i|\} \leq 1/m$. The results then holds on $\left(\bigcup_{m=1}^{\infty} N_m\right)^{\mathrm{c}}$.

Lemma 2: $\forall x \in S$, $\frac{1}{k_n}(1 + \tilde{x}^2)$ is uniformly bounded, with $\frac{1}{k_n}(1 + \tilde{x}^2) \xrightarrow{\mathrm{P}} 0$.

Proof: Chebychev's inequality yields $\tilde{x}^2 \leq k_n$. If x is an atom, then by definition and the law of large numbers, $\tilde{x}^2 \xrightarrow{\text{as}} 0$. Hence without loss of generality it may be assumed x is not an atom. In the terminology of section 2 it is sufficient to show $\tilde{x}^2 \xrightarrow{\mathrm{P}} 3$, or equivalently, $(\hat{\sigma}_n^2 + \hat{\mu}_n^2)/\hat{\mu}_n^2 \xrightarrow{\mathrm{P}} \frac{4}{3}$.

Let

$$X_j^* = \begin{cases} x - X_j + [F(X_j) - F(X_j^-)] \, \eta_j & \text{if } X_j \leq x \\ X_j + [F(X_j) - F(X_j^-)] \, \eta_j & \text{otherwise} \end{cases}$$

where η_j is an independent sequence of uniform random variables. Define $F^*(t) = P(X_j^* \leq t)$. Then F^* has a density bounded away from zero on its support. Further, for $j \in J_n(x)$, $X_j^* \geq x - X_j$.

From hypothesis $\exists r > 0$ such that F is continuous on $[x - r, x]$. Denote the density of F^* by f^*. Pick $\epsilon > 0$. Then $\exists \delta \in (0, r)$ such that $M < (1 + \epsilon)m$, where $M = \max_{z \in [0,\delta)} f^*(z)$, and $m = \min_{z \in [0,\delta)} f^*(z)$. Define

$$G(t) = \begin{cases} F^*(t), & \text{if } t \leq \delta; \\ F^*(\delta) + f^*(\delta)(t - \delta) & \delta < t < \delta + \frac{1}{f^*(\delta)}[1 - F^*(\delta)]; \\ 1 & \text{otherwise.} \end{cases}$$

and set $Z_j = G^{-1} \circ F^*(X_j^*)$. Z_j has cumulative distribution function G, $m \leq \frac{\mathrm{d}}{\mathrm{dx}} G(x) \leq M$, and $Z_j = x - X_j$ for $X_j \in (x - \delta, x]$. Let $Z_{(j)}$ denote the order statistics. From Lemma 1 $\mathrm{P}(\{X_k\}_{k \in J_n(x)} \neq \{x - Z_{(i)}\}_{i=1}^{k_n}) \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$. Thus it suffices to show

$$\frac{\frac{1}{k_n} \sum_{j=1}^{k_n} Z_{(j)}^2}{\left(\frac{1}{k_n} \sum_{j=1}^{k_n} Z_{(j)}\right)^2} \xrightarrow{\mathrm{P}} \frac{4}{3}.$$

Set $U_{(i)} = G(Z_{(i)})$. Then $\{U_{(i)}\}$ are distributed as the order statistics from a uniform distribution. From construction of the Z's and Taylor's theorem, $U_{(i)} = G(Z_{(i)}) = G(0) + Z_{(i)} \frac{\mathrm{d}}{\mathrm{dx}} G(x)|_{x=\eta} = l \cdot Z_{(i)}$, where $\eta \in [0, Z_{(i)}]$ and $m \leq \frac{\mathrm{d}}{\mathrm{dx}} G(x)|_{x=\eta} = l \leq M$. To obtain an upper bound of the limit one applies simple algebra to yield

$$\begin{aligned} \frac{\frac{1}{k_n} \sum_{j=1}^{k_n} Z_{(j)}^2}{\left(\frac{1}{k_n} \sum_{j=1}^{k_n} Z_{(j)}\right)^2} &< \left(\frac{M}{m}\right)^2 \frac{\frac{1}{k_n} \sum_{j=1}^{k_n} U_{(j)}^2}{\left(\frac{1}{k_n} \sum_{j=1}^{k_n} U_{(j)}\right)^2} \\ &= (1 + \epsilon)^2 \frac{\frac{1}{k_n} \sum_{j=1}^{k_n} \{U_{(j)}/U_{(k_n+1)}\}^2}{\{\frac{1}{k_n} \sum_{j=1}^{k_n} U_{(j)}/U_{(k_n+1)}\}^2} \\ &\xrightarrow{\mathrm{P}} \frac{4}{3} (1 + \epsilon)^2 \qquad (a) \end{aligned}$$

Since the random variables $\{U_{(j)}/U_{(k_n+1)}\}$ are distributed according to the ordered statistics of a uniform distribution (a) follows from a simple variant of the weak law of large numbers. A symmetric argument yields the lower bound. Since ϵ is arbitrary the result follows.

Lemma 3: $\forall x \in S,\quad \mathrm{E}\sum_{i \in J}\left(m(x) - m(X_i)\right)^2 = \mathrm{O}(k_n^3/n^2)$.

Proof: Keep the same notation and construction of Lemma 2. Define $W_j = F^*(X_j^*)$. Then W_j are uniformly distributed and from the mean value theorem, $W_j = X_j^* \; f^*(\eta)$ for some $\eta \in [0, X_j^*]$. Let $\gamma = \inf_x f^*(x)$. From hypothesis III $\exists M(x)$ such that $|m(x) - m(X_{(j)})| \leq M(x)\;|x - X_{(j)}|$. Thus

$$\begin{aligned}
\mathrm{E}\sum_{i \in J}\left(m(x) - m(X_j)\right)^2 &\leq M^2(x)\mathrm{E}\sum_{i \in J}(x - X_j)^2 \\
&\leq M^2(x)\mathrm{E}\sum_{i \in J}(X_j^*)^2 \\
&\leq \frac{M^2(x)}{\gamma^2}\mathrm{E}\sum_{j=1}^{k_n} W_{(j)}^2 \\
&\leq \frac{M^2(x)}{\gamma^2}\;\frac{k_n^3}{n^2} \qquad (a)
\end{aligned}$$

Step (a) is a consequence of the fact that $W_{(j)}$ is the j^{th} order statistic from n uniformly distributed random variables and consequently is distributed according to a Beta distribution $B(j, n+1-j)$.

6. Acknowledgements.

The authors appreciate the helpful comments of Iain Johnstone and Fred Huffer.

7. References.

Breiman, L. and Friedman, J.H. (1982) *Estimating Optimal Transformations for Multiple Regression and Correlation*, Stanford University Department of Statistics, Technical Report ORION 010

Breiman, L. and Friedman, J.H. (1984) *Estimating Optimal Transformations for Multiple Regression and Correlation*, J.A.S.A. (to appear)

Friedman, J.H. and Stuetzle, W. (1981) *Projection Pursuit Regression*, J.A.S.A. 76

Friedman, J.H. and Stuetzle, W. (1982) *Smoothing of Scatterplots*, Stanford University Department of Statistics, Technical Report ORION 003

McDonald, J.A. and Owen, A.B. (1984) *Smoothing with Split Linear Fits*, Stanford University Department of Statistics, Technical Report LCS 007

Stone, C.J. (1977) *Consistent Nonparametric Regression*, The Annals of Statistics Vol. 5

Stone, C.J. (1982) *Optimal global rates of convergence for nonparametric regression*, The Annals of Statistics Vol. 10